MEMO (Version 2)

To: Group 2 Consortium

From: K. McLaughlin (scatter@cmr.gov), I. Bóndár

Date: August 20, 2001

Subject: What is the expected deterioration?

This memo attempts to put into context, inherent limitations of a reference event set to validate (or demonstrate) relocation improvement/deterioration for a new set of travel times or calibrations. A "relocation test data set" contains 1) a set of reference origins and 2) a set of associated arrivals. Each reference epicenter has a finite uncertainty, X, that we characterize as GTX. Each arrival has a "measurement error". The travel-time tables for the new model we wish to test are supplied with "model error". The location program combines the *a-priori* "model errors" and "measurement errors" to estimate a 90% confidence coverage ellipse (major axis, minor axis, and strike).

Is it meaningful to improve/degrade the location of an individual GT5 event by 6km? Is it meaningful to improve/degrade any individual location by 6km when the confidence ellipse is 6000 km^2? Individual location improvements/degradations are rarely meaningful on their own. Given that the error budgets are non-zero, we must evaluate calibrations based on sample statistics of a test data set. How do we know if the test data set is large enough or of sufficient quality?

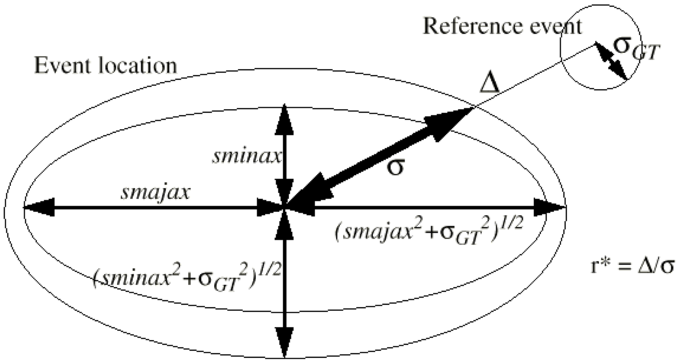

Figure 1. The location confidence ellipse is increased by the uncertainty in the reference event location (GT). The miss location vector normalized by the confidence ellipse in the direction of the reference event called "coverage". Coverage is >1 if GT lies outside the ellipse and < 1 if GT lies inside the ellipse.

In a previous memo, Bóndár suggested that for the purposes of evaluation, we should combine the GTX error with the location confidence ellipse. The resulting "test confidence ellipse" is used to normalize the length of the miss location vector to produce a statistic called "test coverage". In this memo we distinguish the new statistic from the usual “coverage” statistic which does not take into account the GTX uncertainty. Coverage statistics have following properties. If the GT epicenter is within the coverage ellipse, then coverage is less than 1. If the GT epicenter is outside the coverage ellipse, then coverage is greater than 1. If the total error model is correct then we expect coverage should be distributed Chi-squared with 2 degrees of freedom, the median value (50^{th} percentile) should be about 0.3 and the 90^{th} percentile 1.0.

In this memo, we carry these concepts a step further to ask the question, “Given a reference event test set, and finite errors, how many events should we expect to deteriorate?” Alternatively we might ask, “What are the limitations of the test data set?” Since GTX errors, measurement errors, and model errors are not zero, we should expect some fraction of the events to deteriorate.

If we center the “test confidence ellipse” on the reference event (GT) location, then we can measure the miss location of both the new location (calibrated) as well as the old location (uncalibrated IASEPEI) in the same coordinate system. In this way, we define the normalized miss location of the old location as E1 and the normalized miss location of the new location as E2. E2 is identical to the “test coverage statistic” for the new location. We can interpret E1< 1 or > 1 as to whether the old location with the new error model would have “covered” GT location.

Consider the following Monte Carlo experiment. Begin with the GT location and for realization, generate synthetic arrivals with Gaussian errors consistent with the total error model (GTX + measurement + model) and locate the event using the synthetic arrivals. Repeat the process for many realizations. We expect 90% of the synthetic locations to lie within the test confidence ellipse. Whether we used the old (uncalibrated) or new (calibrated) travel times is irrelevant to the Monte Carlo results so long as a consistent set is used for both the synthetic arrivals and location estimation.

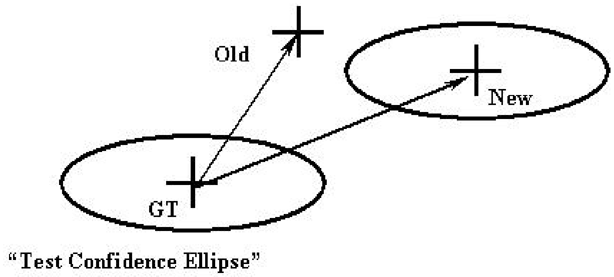

Figure 2. We normalize the old, Δ_1, and new, Δ_2, miss locations by the “test confidence ellipse” centered on the GT location. $E_1 = \Delta_1/\sigma_1, E_2 = \Delta_2/\sigma_2$.

Coverage is usually analyzed at the 90^{th} percentile where the expected value is unity. Table 1 shows the fractions of events for which E1 and E2 are > or < 1 for a test data set of 571 reference events (GT0-10). We located all events using Pn and Sn arrivals (3 to 20 degrees) using the IDC IASPEI travel times (OLD - uncalibrated) and then we relocated the events using a set of SSSCs (NEW - calibrated). We see that 91% of new locations cover the GT location. Therefore, we can conclude the error model predicts “honest” 90% confidence ellipses; E2 < 1 for 90% of the calibrated locations. However, we also see that E1 < 1 for 90% of the uncalibrated locations; the old locations are already within the 90% tolerance of GT uncertainty and new error budget. This is remarkable, given that the new calibrated model standard deviations are about 50% of the old uncalibrated model standard deviations and the overall miss location was reduced from the old model to the new model. 60% of the locations were improved, 47% were improved by more than 20%, and the 80th percentile miss location was reduced by 33%. However, the median miss location was reduced by only 14% and 31% were deteriorated by more than 20%. Obviously, if the GT uncertainty, measurement error, and model error is not zero, then we should expect some events will get worse (deteriorate). However, in order to determine if this number is significant, we need to predict what the distribution should look like. Only when we have answered this question, can we determine if a significant number of events got worse given the error models.

Table 1. E1 uncalibrated "test coverage" (without SSSCs), E2 calibrated "test coverage" (with SSSCs) based on the calibrated 90% test confidence ellipse (E=1) centered on the GT location. When E1 > 1 and E2 > 1 both locations lie outside the 90% test confidence ellipse. When E1 < 1 and E2 > 1 the location moved from inside to outside (worse). When E1 > 1 and E2 < 1 the location moved from outside the ellipse to inside (better). When E1 < 1 and E2 < 1 then both the locations lie inside the ellipse.*

E* = 1.0	**E1 > E***	**E1 < E***	**All**
E2 > E*	13 (f1 = 0.02)	40 (f2 = 0.07)	53 (f1+f2 = 0.09)
E2 < E*	44 (f3 = 0.08)	474 (f4 = 0.83)	518 (f3+f4 = 0.91)
All	57 (f1+f3 = 0.10)	514 (f2+f4 = 0.90)	571 Total

Examination of Table 1 shows the number of locations that moved from inside to outside the test ellipsoid (E1 < 1, E2 > 1) is 7%. These events got worse at the 90% confidence level. This is just slightly less than what we would expect if we lost coverage on 9% of the 90% of events that already had coverage (9% of 90% is 8%). Given the sample size of 571 events, we can expect to see ± 1.3 % fluctuations of the 10th and 90th percentiles. Therefore, the number of events that got significantly worse is about what we would expect by random chance. The important thing to take away from this part of the analysis is that at any test coverage level, we should expect to see a fraction of events get worse, we can define a significance test, and we can predict the fraction of events that could get worse simply by random chance. If the new error model is honest and the old model performs even moderately well then we should expect to observe a level of degradation. The power of the reference test set increases as f3+f4 becomes larger than f2+f4. In the case of Table 1, we can see that f3+f4 = 0.91 is only marginally larger than f2+f4 = 0.9. Therefore this data set has almost no power at a coverage level of 1.

We repeat the analysis for the test confidence ellipse with E* = 0.3 and present the results in Table 2. 56% of the events are within the ellipse calibrated or uncalibrated. 18% of the events are outside the ellipse calibrated or uncalibrated. F2 (8%) is significantly smaller than f3 (17%); 99 events moved from outside to inside while only 46 events moved from inside to outside. Given that we expect f2 and f3 to fluctuate by ~2% this is statistically significant. For the E* = 0.3 test confidence ellipse, f3+f4 = 0.74 is measurably larger than f2+f4 = 0.64 and we are able to detect the difference between the two models. The percentage of events that degraded (8%) is actually only 50% of what we would have expected (26% of 64% = %16) by random chance.

Table 2. E1 uncalibrated "test coverage" (without SSSCs), E2 calibrated "test coverage" (with SSSCs) based on the calibrated 50% test confidence ellipse (E=0.3) centered on the GT location. When E1 > = 0.3 and E2 > 0.3 both locations lie outside the 50% test confidence ellipse. When E1 < 0.3 and E2 > 0.3 the location moved from inside to outside (worse). When E1 > 0.3 and E2 < 0.3 the location moved from outside the ellipse to inside (better). When E1 < 0.3 and E2 < 0.3 then both the locations lie inside the ellipse.*

E* = 0.3	**E1 > E***	**E1 < E***	**All**
E2 > E*	105 (f1 = 0.18)	46 (f2 = 0.08)	151 (f1+f2 = 0.26)
E2 < E*	99 (f3 = 0.17)	321 (f4 = 0.56)	518 (f3+f4 = 0.74)
All	204 (f1+f3 = 0.36)	367 (f2+f4 = 0.64)	571 Total

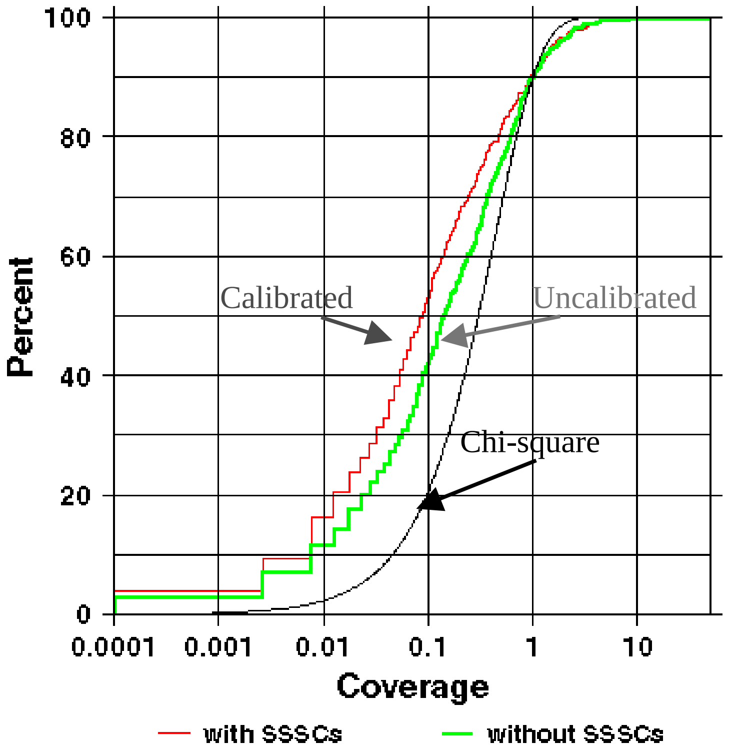

Figure 3. Plots of cumulative E1 (uncalibrated without SSSCs) and E2 (calibrated with SSSCs) test coverage for the GT0-10 reference event test data set. The expected Chi-squared distribution is shown for reference. Both the calibrated and uncalibrated travel-time tables perform better than expected for 90% of the events based on the theoretical Chi-squared distribution. Both calibrated and uncalibrated travel-time tables perform poorer than expected for 10% of the events based on the theoretical Chi-squared distribution. For test values of E between 0.01 and 0.6, between 50% and 100% more*

events move from outside to inside as apposed to move from inside to outside the test ellipse with calibration.

Figure 3 shows the cumulative distributions of E1 and E2 for the uncalibrated (without SSSCs) and calibrated (with SSSCs) test coverages. We see that there is no measurable difference between the two models for E1 and E2 greater than 1. However, at all values less than 1, the calibrated model performs measurably better. The difference between the two cumulative plots is (f2+f4) – (f1+f2) as a function of the test coverage value, E*. If we repeat the test for values of E* between 0.01 and 0.6 we find that 50 to 100% more events move from outside to inside as move from inside to outside the test ellipse depending upon the chosen value of E*. Just as important, for all values of E* < 1, the number of events that degrade is never larger than what we would expect from random chance.

We define δ = E1 – E2 as our "normalized improvement". Using normalized improvement, we can identify those events that are significantly improved versus those that are significantly degraded. Table 3 shows some of the relevant statistics. If we choose a test value E* = 0.3, then we observe 12% were Degraded, 20% were Improved, and 68% were neither. The ratio of Improved/Degraded is a maximum near the test value E* = 0.5 where 1.85 times more events are Improved than Degraded.

For E* = 0.5, 77% of the events are neither Improved nor Degraded. For a test value of E* = 0.5 conditioned on the test data set we would expect 14% (19% of 76%) of the events would move from inside to outside the test ellipse with calibration (got worse) compared to 8% in Table 4. Table 4 shows that for E* = 0.5, 15% of the events moved from outside to inside (got better) which is almost what we would expect (24% of 81%).

Table 3. Numbers and fractions of events with normalized improvement, δ = E1 – E2, > +E or < -E* for test values of E*. For a given test value, δ > +E* is "Improved" and δ < -E* is "Degraded".*

E*	**δ < -E* (Degraded)**		**δ > +E* (Improved)**		**Improved/Degraded**
0.0	242	42%	329	58%	1.36
0.1	114	20%	192	34%	1.68
0.2	84	15%	138	24%	1.64
0.3	70	12%	115	20%	1.64
0.4	56	10%	96	17%	1.71
0.5	46	8%	85	15%	1.85
0.6	43	8%	70	12%	1.63
0.7	42	7%	57	10%	1.36
0.8	33	6%	39	7%	1.18
0.9	29	5%	29	5%	1.00
1.0	26	5%	26	5%	1.00
2.0	9	2%	11	2%	1.22

Table 4. E1 uncalibrated "test coverage" (without SSSCs), E2 calibrated "test coverage" (with SSSCs) based on E=0.5.*

E* = 0.5	**E1 > E***	**E1 < E***	**All**
E2 > E*	53 (f1 = 0.09)	53 (f2 = 0.09)	106 (f1+f2 = 0.19)
E2 < E*	83 (f3 = 0.15)	382 (f4 = 0.67)	518 (f3+f4 = 0.81)
All	136 (f1+f3 = 0.24)	435 (f2+f4 = 0.76)	571 Total

In conclusion, we propose new statistical tests for evaluating calibration performance in the presence of uncertainties in reference GT accuracy, measurement, and model errors. Individual location improvements/degradations are rarely meaningful. Therefore, we must evaluate calibrations based on sample statistics of uncertain test data sets. The new set of tests will help evaluate whether the test data sets and the results can be expected to make meaningful statistical statements about calibration performance. Using the test coverage tables defined above it is possible to distinguish between whether a set of calibrations does better or worse than would be expected by random chance. Using the normalized improvement statistic defined above it is possible to test if the relocation of an individual event is significant.